Ruler and Compass Constructions and Abstract Algebra

Introduction

Around 300 BC, Euclid wrote a series of 13 books on geometry and number theory. These books are collectively called the Elements and are some of the most famous books ever written about any subject.

In the Elements, Euclid described several "ruler and compass" constructions. By ruler, we mean a straightedge with no marks at all (so it does not look like the rulers with centimeters or inches that you get at the store). The ruler allows you to draw the (unique) line between two (distinct) given points. The compass allows you to draw a circle with a given point as its center and with radius equal to the distance between two given points.

But there are three famous constructions that the Greeks could not perform using ruler and compass:

- Doubling the cube: constructing a cube having twice the volume of a given cube.
- Trisecting the angle: constructing an angle 1/3 the measure of a given angle.
- Squaring the circle: constructing a square with area equal to that of a given circle.

The Greeks were able to construct several regular polygons, but another famous problem was also beyond their reach:

- Determine which regular polygons are constructible with ruler and compass.

These famous problems were open (unsolved) for 2000 years!

Thanks to the modern tools of abstract algebra, we now know the solutions:

• It is impossible to double the cube, trisect the angle, or square the circle using only ruler (straightedge) and compass.

• We also know precisely which regular polygons can be constructed and which ones cannot.

(Not everyone seems to be aware of this, however. To this day, mathematicians around the world occasionally receive communications from people claiming to have found a method to trisect the angle, for example).

The idea of using algebra to bear on geometry problems, and vice-versa, is very beautiful and powerful. You are already familiar with analytic geometry, where you introduce a system of coordinates (Cartesian or otherwise) that allows you to describe many geometric objects and their measures (lengths, areas, volumes, angle measures) through algebraic equations and manipulations.

We will embark on a road that also uses algebra to solve these famous old problems from classical Greek geometry. But the algebra that we need is more sophisticated. It deals with generalized algebraic structures and their properties, and we call it abstract algebra.

Some of the objects from abstract algebra that we will meet along our journey are groups, rings, fields, rings of polynomials, field extensions, algebraic and transcendental numbers, and vector spaces.

(Vector spaces are also protagonists in the subject called linear algebra, which studies linear transformations between vector spaces).

Overview

• We will start by reviewing some of Euclid's constructions that you learned in high school geometry.

• Next, we will identify the points in the plane with the complex numbers, and we will specify precisely what it means to construct a point (number) with ruler and compass and what it means for a number to be constructible.

• Then we will take a long dive into abstract algebra and introduce the objects mentioned above, together with the relevant facts that we will need.

• Finally, we will use these algebraic tools to describe properties of the constructible numbers and to show why the three famous constructions (doubling the square, trisecting the angle, squaring the circle) are impossible.

• If time allows, we will say a few words (without any technical details) about the solution of the other problem, namely determining precisely which regular polygons can be constructed.

Constructions from High School Geometry

In your high school geometry class, you probably learned several of Euclid's ruler and compass constructions. Here are some that you should know how to do:

(1) Bisect a line segment.

(2) Draw the perpendicular to a given line through a given point on the line.

(3) Draw the perpendicular to a given line through a given point not on the line.

(4) Bisect an angle.

(5) Copy a given angle so its vertex is a given point and one of its rays lies on a given line (and we can also choose the half-plane of this line on which to copy the angle).

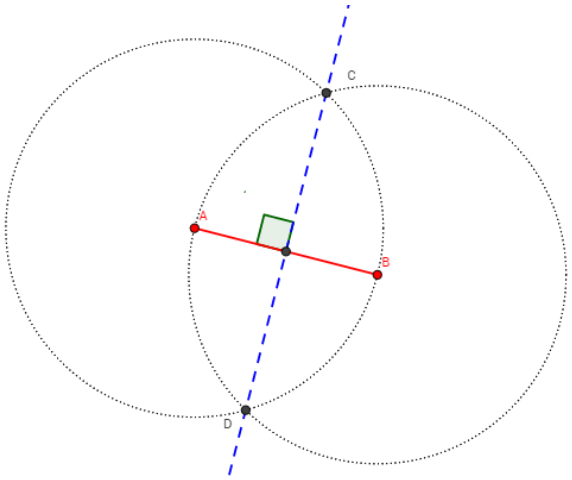
(6) Draw the line parallel to a given line through a given point not on the line (the Parallel Postulate says this line is unique).

(7) Draw a triangle that is similar to a given triangle, with a prescribed side.

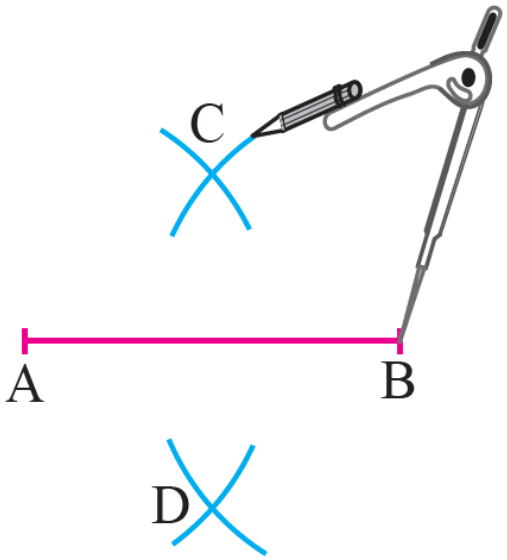
(8) Subdivide a line segment into n equal parts, for any $n \geq 1$.

For example, constructions (1) – (3) all rely on the fact that the perpendicular bisector of a line segment is also the set of all points that are equidistant from its two endpoints. This fact is easily proved by properties of isometries (distance-preserving transformations of the plane, also known as rigid motions) or by triangle congruence criteria.

This fact allows you to construct the perpendicular bisector of a line segment very easily:



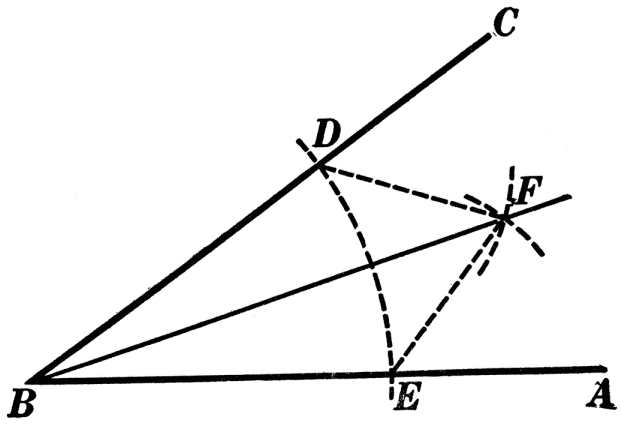
OR:



Therefore, you not only bisected the line segment, which is construction (1), but you also constructed a line perpendicular to a given line.

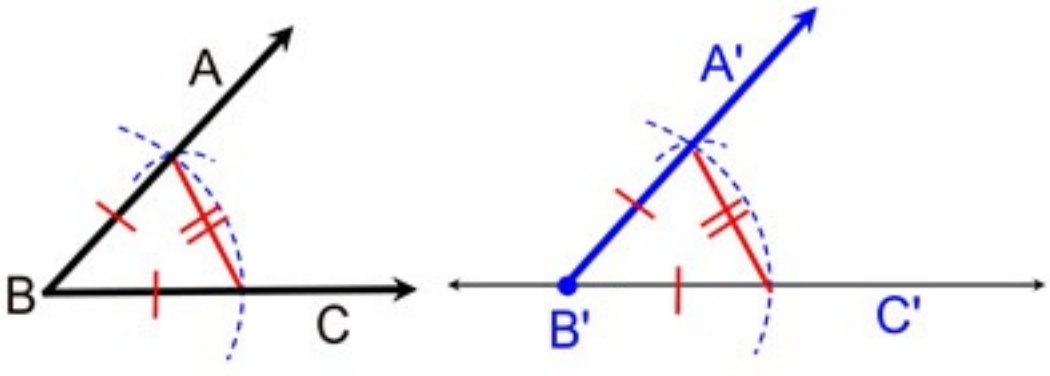
Question 1: Can you modify or add to this construction to perform (2) and (3)?

Construction (4), bisecting an angle, can be accomplished by making use of the SSS triangle congruence criterion:



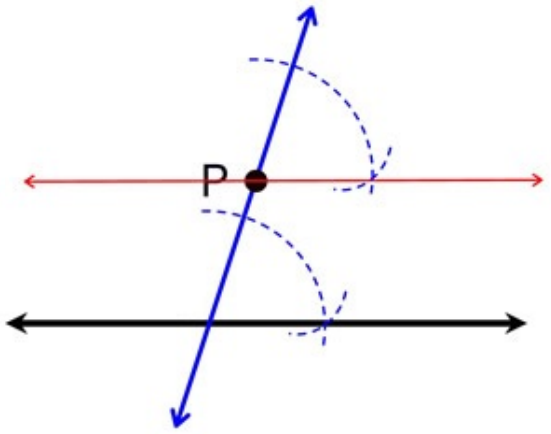
Question 2: Can you prove that the angle bisector of an angle (of measure less that 180^0) is the ray (inside the angle and with initial point at the vertex) whose points are equidistant from the sides of the angle?

Construction (5), copying an angle, also relies on SSS:



Constructions (6) and (7) are both based on construction (5).

The construction of parallel lines (6) relies on the theorem that if a transversal to two lines produces congruent corresponding angles, then the two lines are parallel:



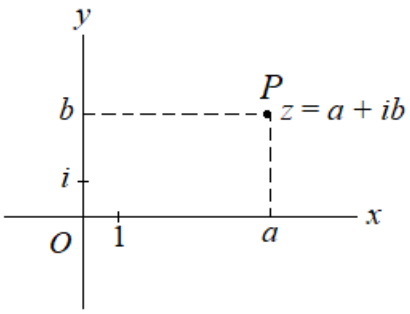
The construction of similar triangles (7) relies on the AA criterion for triangle similarity. It is based on copying two angles whose vertices are the endpoints of a given line segment.

Finally, construction (8) is based on construction (6) of parallel lines and on easy proportions resulting from similar triangles.

Question 3: Explain how to use ruler and compass to subdivide a line segment into n equal parts.

The Set of Constructible Numbers

As you know, we can identify the Euclidean plane with the set $\mathbb{C}$ of all complex numbers: we set up a coordinate system, for example a system of Cartesian coordinates, and then every point P in the plane can be uniquely identified with its coordinates (a, b), or equivalently with the complex number $z = (a, b) = a + ib$:



With this identification, points in the plane are complex numbers to us, and constructing a point is the same as constructing a complex number.

But what exactly does it mean to construct a point or number? We should give a precise definition. We are allowed the following constructions of lines and circles:

R: Given two distinct points z_1, z_2, we can draw the line that passes through them.

For this, of course, we use the ruler.

C: Given a point z_1 and two distinct points z_2, z_3, we can draw the circle with center z_1 and radius $|z_2 - z_3|$.

For this, of course, we use the compass.

From these lines and circles, their points of intersection (when not empty) give us new points as follows:

P_{ll} = the point of intersection of two distinct lines constructed as above.

P_{lc} = the point(s) of intersection of a line and a circle constructed as above.

P_{cc} = the point(s) of intersection of two distinct circles constructed as above.

Now we can use these new points to construct more lines and circles, which in turn yield new points, etc.

Definition: The point $z \in \mathbb{C}$ is constructible if there is a finite sequence of ruler and compass constructions using R, C, P_{ll}, P_{lc} and P_{cc} that begins with 0 and 1 and ends with z.

We define $\mathfrak{C} = \{z \in \mathbb{C} \mid z \text{ is constructible}\}$.

This is the set of all constructible numbers.

Examples:

• Every $m \in \mathbb{Z}$ is constructible ($\mathbb{Z}$ is the set of all the integers).

• Every "Gaussian integer" $m + in \in \mathbb{Z}[i]$ is constructible.

• The points $\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$ are constructible: they are the intersection points of the two circles with radius 1 having centers at 0 and 1, respectively.

• z is constructible if and only if $\overline{z}$ (its complex conjugate) is constructible. This is because of symmetry about the x-axis. Any construction that starts with 0 and 1 and ends with z can be reflected in an obvious way about the x-axis to end with $\overline{z}$. This will also be clear once we establish a few basic facts about constructible numbers (they form a "field", and $z = a + ib$ is constructible if and only if both a and b are constructible).

Question 4: Explain how to construct $z = m + in$ where $m, n \in \mathbb{Z}$.

Example: Since we can copy any given angle on a given initial ray, the regular n-gon is constructible if and only if $\xi_n = e^{2\pi i/n}$ is constructible:

• If the regular n-gon is constructible, then we can construct an angle of measure $2\pi/n$ using the center and two adjacent vertices of the n-gon (if the center is not yet a given point, we can easily construct it, for example, as the intersection of the bisectors of the interior angles of the n-gon). Then we copy this angle so that it has vertex at the origin and one of its rays is the positive x-axis. We can also arrange it so that the other ray lies on the upper half plane. Then the intersection of that other ray with the unit circle is ξ_n.

• If ξ_n is constructible, then we can construct the angle of measure $2\pi/n$ whose vertex is 0 and whose rays go through 1 and ξ_n respectively. Copying this angle repeatedly using always 0 as the vertex, and then intersecting the resulting rays with the unit circle, we obtain all the n^{th} roots of 1, namely $1, \xi_n, \xi_n^2, \ldots, \xi_n^{n-1}$. Joining these points in the obvious way, we obtain a regular n-gon.

Remark # 1: Playing by the Rules

We only allow exact constructions, and, by definition, they must be achieved in a finite number of steps. Also, we do not accept any other instruments except for the unmarked ruler and the compass.

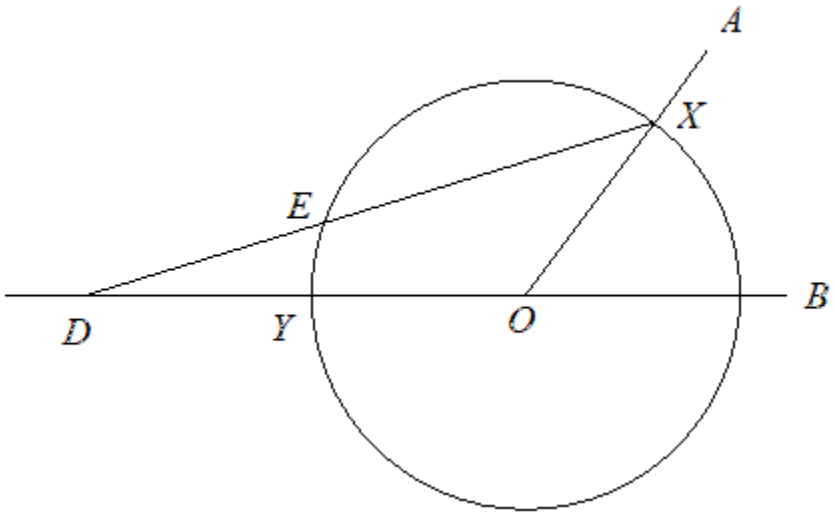
If we ease any of these restrictions, then more constructions become possible. That is, more numbers can be obtained.

Take for example the problem of trisecting the angle.

• If we allow a ruler that has two points marked on it, together with a compass, then we can trisect any angle:

Say the two marked points on the ruler are a distance r apart. Given an angle $\angle AOB$ with measure θ, draw the circle with center O and radius r. Let X be the point where the ray $\overrightarrow{OA}$ meets the circle and let Y be the point where the ray opposite to $\overrightarrow{OB}$ meets the circle.

Now place the ruler with its edge on X and one mark on the ray $\overrightarrow{OY}$ at a point D. Slide it until the other marked point lies on the circle at a point E. Then $\angle EDO$ has measure $\theta/3$:



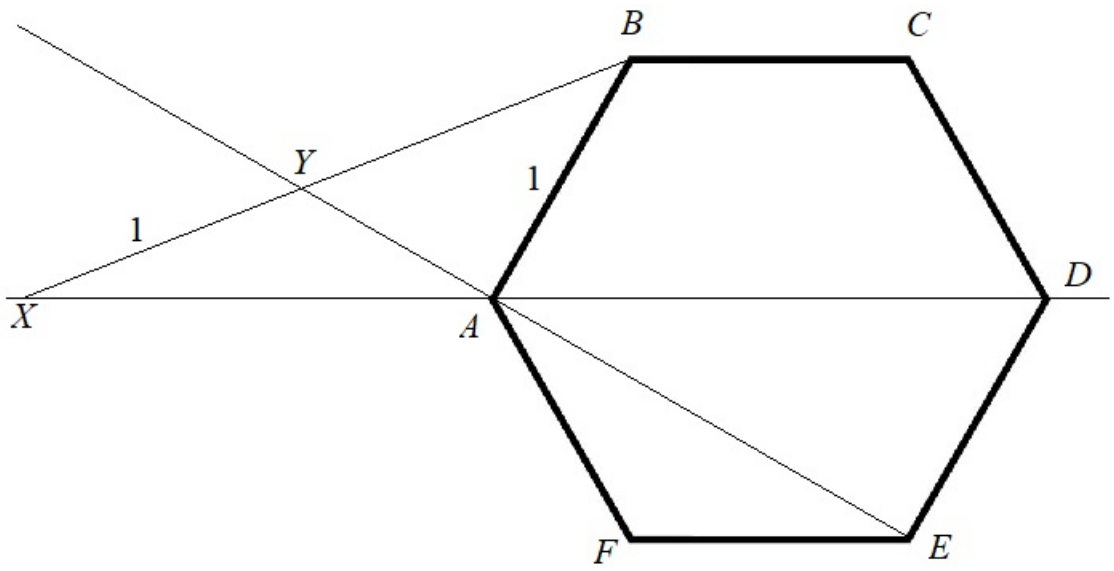
Question 5: Show that $\angle EDO$ has measure $\theta/3$, where θ is the measure of $\angle AOB$.

• With (unmarked) ruler and compass only, we can also approximate the trisection of any angle to any desired (but not exact) accuracy:

The geometric series $\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\cdots$ converges to $\frac{\frac{1}{4}}{1-\frac{1}{4}}=\frac{1}{3}$. Given an angle of measure θ, we can bisect it repeatedly, as many times as we want. So we can find an angle of measure $\frac{\theta}{2^n}$ for any $n \geq 1$. Since we can copy (and therefore add) angles, we can construct an angle of measure as close to $\theta/3$ as we want, by using sufficiently many terms of this geometric sequence.

• If we allow constructions with infinitely many steps, then the above geometric sequence can be used to trisect any angle.

If you have a compass and a ruler with just two marks on it, not only you can trisect the angle (as we just saw), but you can also double the cube:



In the picture, $ABCDEF$ is a regular hexagon with side length 1, and $XY = 1$.

Question 6 (Challenge):

(a) Show that $YB = 2^{1/3} = \sqrt[3]{2}$.

(b) Explain how you can use compass and a marked ruler with two marks at a distance 1 from each other to construct a segment of length of length $\sqrt[3]{2}$.

(c) Explain why the cube can be doubled with compass and a twice-marked ruler.

Remark # 2: Constructions from a Given Set of Points

If, instead of starting our constructions from the initial set of points $\{0, 1\}$, we start from an arbitrary finite set P that contains 0 and 1, then we can talk about statements like "bisecting a given angle" or "bisecting a given segment" without having to worry about whether this given angle or segment is constructible.

This is in fact the more natural setting for the high school constructions that we reviewed above. For example, we can bisect any given segment regardless of its length, because if the segment is given, then we can use its length in our construction. The length of the segment may not be a constructible number: it may be impossible to construct it if we start from 0 and 1. But we do not care because this number is just given to us, so we can open the compass to this length to draw a circle.

The entire theory of ruler and compass constructions works in the same way if we start from 0 and 1 or if we start from a bigger set P of given points, even if the set of constructible points gets bigger.

Having a bigger finite set of initial points does not change the impossibility of trisecting the angle or doubling the cube or squaring the circle, because these problems ask for a method that will trisect *any* angle, or double *any* cube, or square *any* circle. We can obviously trisect some angles (for example, 180^0 or 90^0), and if we start with more given points, we may be able to trisect more angles. But there will always be angles that we cannot trisect.

Overall Strategy: a Look Ahead

From now on, things will get a little more abstract and sometimes technical. So it is a good idea to keep in mind our objective and an overall strategy to get there.

• We want to prove that certain constructions are impossible to achieve with (unmarked) ruler and compass.

• It is easy to see that this is equivalent to proving that certain numbers are not constructible.

• Using tools of abstract algebra, we will eventually find a property that a number must possess in order for it to be constructible.

• Finally, we will show that certain numbers do not possess this property, and therefore are not constructible, and therefore certain constructions are impossible to achieve by ruler and compass.

We need to develop the algebraic tools needed to find that useful property that all constructible numbers must possess, and this will take a significant amount of work. But as a reward, the concepts from abstract algebra that we will discuss are of paramount importance in all branches of higher mathematics.

Groups

Recall that a binary operation on a set S is a function $*: S \times S \to S$. We use the notation $*(a, b) = a * b$.

Definition: A group is a set G and a binary operation $*$ on G such that

• Associativity: $(a * b) * c = a * (b * c) \quad \forall\, a, b, c \in G$

• Identity: $\exists\, e \in G$ such that $e * g = g * e = g \quad \forall\, g \in G$

• Inverses: $\forall\, g \in G,\ \exists\, g^{-1} \in G$ such that $g * g^{-1} = g^{-1} * g = e$

Examples:

• $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$ (the sets of all integers, rational numbers, real numbers, and complex numbers respectively) are all groups under the operation of addition.

• However, none of these sets is a group under multiplication. **Why?**

• If we remove 0 from each of these sets, we get the sets $\mathbb{Z}^*, \mathbb{Q}^*, \mathbb{R}^*$ and $\mathbb{C}^*$. Three of them are groups under multiplication, and one is not. **Which one is not?**

• Let $n > 1$ be an integer. The set $\mathbb{Z}_n = \{0, 1, 2, \dots, n-1\}$ under addition modulo n is a group.

• The set $M_n(\mathbb{R})$ of $n \times n$ matrices with real entries is not a group under matrix multiplication, but its subset consisting of all invertible matrices *is* a group.

Example: A very important example of a group is the following. Remember that a function is a bijection if it is both injective (one-to-one) and surjective (onto).

Let A be any nonempty set. The set of all bijections $A \to A$, which are called permutations of A, forms a group under the operation of composition of functions. The identity element is the identity function on A, and the inverse of a bijection is its inverse as a function. This group is usually denoted by S_A.

In particular, if A is the set $\{1, 2, 3, \dots, n\}$, then the resulting group is called the group of permutations on n letters, or more commonly, the symmetric group on n letters, and we denote this group by S_n.

The permutations of a mathematical object that preserve some feature(s) is a group, and groups of this kind arise in many branches of mathematics.

For example, the subset of all permutations of the plane that preserve the distance between points is a group, because composition of functions preserves this feature, as do inverse functions. This is the group of all isometries or rigid motions of the plane.

Definition: if the operation of a group G is commutative, i.e., $ab = ba \quad \forall a, b \in G$, then we say that G is abelian.

Examples:

• The group of permutations of a set (having more than 2 elements) is not abelian, because composition of functions is not commutative.

• The group of invertible $n \times n$ matrices under matrix multiplication is not abelian either, because matrix multiplication is not commutative.

• The other groups that we mentioned above are all abelian, because addition and multiplication of complex numbers is commutative, and so is addition modulo n.

Question 7: Prove that in any group G,

(a) The identity element $e \in G$ is unique.

(b) Given $g \in G$, g^{-1} is unique.

Hint: The standard way to prove the uniqueness of an object having some property is to show that if A and B are two such objects, then $A = B$.

For part (a), show that if e and e' are both identity elements of G, then $e = e'$.

For part (b), show that if h and h' are both inverse elements of g, then $h = h'$.

Rings and Fields

For us, a ring will always mean a "commutative ring with unity". This means that multiplication is commutative and there is a multiplicative identity. If you take a class in abstract algebra, you will see a more general definition.

Definition: A ring is a set R with two binary operations, + (addition) and · (multiplication), such that

• R is an abelian group under addition. We denote the additive identity by 0.

• Multiplication is associative: $(ab)c = a(bc) \quad \forall a, b, c \in R$

• Multiplication is commutative: $ab = ba \quad \forall a, b \in R$

• Multiplicative identity: $\exists\ 1 \in R$ such that $1a = a \quad \forall a \in R$

• The distributive law holds: $a(b + c) = ab + ac \quad \forall a, b, c \in R$

Note that we usually denote $a \cdot b$ simply by ab.

Examples:

• $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all rings (with the usual addition and multiplication).

• $\mathbb{Z}_n$ is a ring with addition and multiplication modulo n.

• If R is a ring, the set of all polynomials with coefficients in R is denoted by $R[t]$. This is also a ring. We will be very interested in this kind of ring, especially when R is a field (which we will define below).

Question 8: Prove that if R is a ring, then $0a = 0 \quad \forall a \in R$.

Two nonzero elements a, b of a ring such that $ab = 0$ are called zero divisors.

A ring with no zero divisors is called an integral domain. So in an integral domain, $ab = 0 \Longrightarrow a = 0$ or $b = 0$ (the converse is true in any ring by Question 8).

A unit in a ring is an element that has a multiplicative inverse. That is, it is an element u for which there is an element v such that $uv = 1$.

We denote by $-a$ the additive inverse of an element a and by u^{-1} or $\frac{1}{u}$ the multiplicative inverse of an element u.

Question 9:

(a) Prove that $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$ are integral domains.

(b) Is $\mathbb{Z}_5$ an integral domain? How about $\mathbb{Z}_6$? Explain your answers.

Definition: A field is a ring with $1 \neq 0$ whose nonzero elements are all units. That is, every nonzero element has a multiplicative inverse.

If K is a field, then the set $K^* = K - \{0\}$ of all nonzero elements is an abelian group under multiplication.

Question 10: Prove that every field is an integral domain.

Examples:

• $\mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$ are fields, but $\mathbb{Z}$ is not.

• If p is prime, then $\mathbb{Z}_p$ is a field. **Can you prove this?** But if n is composite, then $\mathbb{Z}_n$ is not a field. In fact, it is not even an integral domain. **Can you prove this?**

The Complex Numbers

Since we identified the Euclidean plane with the field $\mathbb{C}$ of complex numbers, this is the stage in which we will work for ruler and compass constructions. Here are some things you should remember about $\mathbb{C}$:

• How to perform arithmetic in $\mathbb{C}$, both in rectangular and in polar coordinates. You should know the geometric properties of the arithmetic operations. Remember that $e^{i\theta} = \cos\theta + i\sin\theta$.

• The absolute value or magnitude $|z|$ of a complex number z. If $z = a + ib = (a, b)$, then $|z| = \sqrt{a^2 + b^2} \in \mathbb{R}$ is the distance from z to the origin O.

• Conjugation: if $z = a + ib = (a, b)$, then its conjugate is $\bar{z} = a - ib = (a, -b)$. Remember that $z\bar{z} = |z|^2$, and therefore $z^{-1} = \frac{\bar{z}}{|z|^2}$ for $z \neq 0$.

Also, $\overline{z_1 + z_2} = \bar{z_1} + \bar{z_2}$ and $\overline{z_1 z_2} = \bar{z_1}\bar{z_2}$.

• n^{th} roots: every $0 \neq z \in \mathbb{C}$ has exactly n distinct n^{th} roots, located at the vertices of a regular n-gon centered at O. If $0 \neq z = re^{i\theta}$, let $r^{1/n}$ be the unique positive (that implies real, of course) n^{th} root of $r > 0$ and let $\xi_n = e^{i\frac{2\pi}{n}}$. Then the n^{th} roots of z are $r^{1/n}e^{i\frac{\theta}{n}} = w, w\xi_n, w\xi_n^2, \ldots, w\xi_n^{n-1}$.

• In particular, the n^{th} roots of 1 are $1, \xi_n, \xi_n^2, \ldots, \xi_n^{n-1}$.

Finally, the complex numbers have a formidable property: $\mathbb{C}$ is algebraically closed. This fact is also called the Fundamental Theorem of Algebra (FTA), and it means that every non-constant polynomial with coefficients in $\mathbb{C}$ has a root, or zero, in $\mathbb{C}$.

An immediate consequence of FTA is that every non-constant polynomial with complex coefficients factors into linear factors over $\mathbb{C}$. Another immediate consequence is that a polynomial of degree n with complex coefficients has exactly n zeros, counting multiplicity.

Note that $\mathbb{Q}$ and $\mathbb{R}$ are not algebraically closed. For example, the polynomial $t^2 - 2$ has rational coefficients but no rational roots. The polynomial $t^2 + 1$ has real coefficients but no real roots. In $\mathbb{C}$, any (non-constant) polynomial has a zero, no matter how large its degree.

The Ring of Polynomials

If R is a ring, the ring of polynomials over R in the indeterminate t is

$$R[t] = \{p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n \mid n \in \mathbb{Z}^{\geq 0},\ a_i \in R\}$$

The a_i are called the coefficients of $p(t)$. We say that $p(t) \in R[t]$ is a polynomial over R, or with coefficients in R.

Sometimes we will simply write p instead of $p(t)$, when the indeterminate is understood.

Example: The polynomial $p = p(t) = 3t^5 - 7t^3 + t^2 + 2t - 3$ is a polynomial over $\mathbb{Z}$, since its coefficients are all integers. That is, $p \in \mathbb{Z}[t]$.

Since $\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$, we also have $p \in \mathbb{Q}[t],\ p \in \mathbb{R}[t],\ p \in \mathbb{C}[t]$.

The usual polynomial addition and multiplication make $R[t]$ a ring. The additive identity is 0, the zero polynomial, and the multiplicative identity is the constant polynomial 1.

We will be mostly interested in the ring $K[t]$ where K is a field. Note that $K[t]$ is an integral domain but not a field. **Why?**

Definition: Let $0 \neq p \in R[t]$. If $p = a_0 + a_1 t + \cdots + a_n t^n$ where $a_n \neq 0$, then the degree of p is n. We write $\partial p = n$.

We can either say that the zero polynomial has no degree, or we can define $\partial 0 = -\infty$, with the symbol $-\infty$ obeying the rules $-\infty < n, -\infty + n = -\infty$ $\forall n \in \mathbb{Z}$ and $(-\infty) + (-\infty) = -\infty$. Then we can easily see that

Claim: If $p, q \in R[t]$, then $\partial(p + q) \leq \max\{\partial p, \partial q\}$, and if R has no zero divisors, then $\partial(pq) = \partial p + \partial q$.

A constant polynomial is a polynomial of degree 0, or the zero polynomial. A linear polynomial is a polynomial of degree 1. A quadratic, cubic, quartic, and quintic polynomial is a polynomial of degree 2, 3, 4, and 5 respectively.

Divisibility in $K[t]$

Let K be any field. If you want, you can just think that $K = \mathbb{C}$.

Division Algorithm: Let $f, g \in K[t]$ with $f \neq 0$. There exist unique polynomials $q, r \in K[t]$ such that $g = fq + r$ and $\partial r < \partial f$.

(q is the quotient and r is the remainder when you divide g by f).

The Division Algorithm in $K[t]$ works just like in $\mathbb{Z}$, except that instead of requiring the remainder to be smaller than the divisor, we require it to have degree smaller than that of the divisor.

Division of one polynomial by another also works almost exactly as division with remainder in $\mathbb{Z}$.

Question 11: Find the quotient and remainder when dividing $g = t^7 - t^3 + 5$ by $f = t^3 + 7$.

Definition: Let $f, g \in K[t]$. We say that f divides g, or f is a factor of g, or f is a divisor of g, or g is a multiple of f, if $\exists h \in K[t]$ such that $g = fh$.

We use the notation $f|g$ when f divides g and $f \nmid g$ when f does not divide g.

An important consequence of the Division Algorithm is the Factor Theorem:

Factor Theorem: Let $p(t) \in K[t]$ with $\partial p > 0$, and let $\alpha \in K$.

Then $p(\alpha) = 0$ if and only if $(t - \alpha)|p(t)$ in $K[t]$.

Question 12: Prove the Factor Theorem.

Hint: Apply the Division Algorithm with $g = p(t)$ and $f = t - \alpha$.

The Factor Theorem is the reason that FTA implies that every non-constant polynomial over $\mathbb{C}$ factor completely into linear factors.

Definition: Let $f, g \in K[t]$. We say that $d \in K[t]$ is a highest common factor (hcf) or greatest common divisor (gcd) of f and g if

- $d|f$ and $d|g$
- $(e|f$ and $e|g) \Longrightarrow e|d$

We use the notation $\text{hcf}(f, g)$, or $\gcd(f, g)$, or sometimes simply (f, g).

$\text{hcf}(f, g)$ is not unique, but it is "almost unique": it is unique up to a constant multiple. We will usually say "the hcf of f and g" with the assumption that this is understood.

Exactly as it happens in $\mathbb{Z}$, the Division Algorithm is the basis for another algorithm that produces $\text{hcf}(f, g)$ for nonzero polynomials f and g over K. It is called the Euclidean Algorithm. It is based on repeated applications of the Division Algorithm and a very simple but important fact (that also works just like in $\mathbb{Z}$):

$$\text{If } a|b \text{ and } a|c \text{ in } K[t], \text{ then } a|(pb + qc) \quad \forall p, q \in K[t]$$

The Euclidean Algorithm not only produces $\text{hcf}(f, g)$, but also allows you to find a combination of f and g that equals $\text{hcf}(f, g)$. That is, it allows you to explicitly find polynomials a and b as in the following lemma:

Lemma: Let $f, g \in K[t]$ be nonzero polynomials and let $d = \text{hcf}(f, g)$. Then $\exists a, b \in K[t]$ such that $d = af + bg$.

Irreducible Polynomials and Unique Factorization

Definition: Let R be a ring and let $f \in R[t]$ be a nonconstant polynomial. We say that f is irreducible in $R[t]$, or irreducible over R, if it cannot be written as a product $f = gh$ of polynomials over R with $\partial g < \partial f$ and $\partial h < \partial f$. Equivalently, if $f = gh$, then g or h is constant.

Of course, if f can be written as a product of two polynomials over R, both of lower degree, then we say that f is reducible over R.

Notes:

• Let $R_1 \subseteq R_2$ and $f \in R_1[t]$. Obviously, if f is reducible over R_1, then it is reducible over R_2. So if f is irreducible over R_2, then it is irreducible over R_1.

The converse is false, of course. For example, $t^2 - 2$ is irreducible over $\mathbb{Z}$ or $\mathbb{Q}$ but reducible over $\mathbb{R}$. [But see Gauss's Lemma below for an important exception: if a polynomial is irreducible over $\mathbb{Z}$, then it is also irreducible over $\mathbb{Q}$].

• Polynomials of degree 1 are always irreducible.

• If $f \in K[t]$ has degree 2 or 3, then it is clear from the Factor Theorem that f is reducible over K (a field) if and only if it has a zero in K. Equivalently, if f has no zeros in K, then if is irreducible over K. Of course, this is not true if $\partial f > 3$.

Definition: $f, g \in K[t]$ are coprime if $\text{hcf}(f, g) = 1$.

We also say that f is prime to g.

Lemma: Let f be irreducible in $K[t]$. If $f|gh$, then $f|g$ or $f|h$.

(This lemma says that in $K[t]$, an irreducible element is prime. The converse is true in any integral domain).

The ring of polynomials over a field, $K[t]$, has the important property of being a unique factorization domain (UFD), like $\mathbb{Z}$:

Theorem: Factorization into irreducible polynomials in $K[t]$ is unique (up to the order of the factors and a constant multiple).

Irreducibility Criteria

It is useful to know if a polynomial is irreducible. Not only we do not need to bother looking for factors (an irreducible polynomial does not have nontrivial factors), but the zeros and the degree of an irreducible polynomial will play a crucial role later.

In general, deciding if a polynomial over a ring is irreducible is not easy (of course, there are exceptions. For example, since $\mathbb{C}$ is algebraically closed (FTA), the only irreducible polynomials over $\mathbb{C}$ are the linear polynomials).

We will be mainly interested in the irreducibility of polynomials over $\mathbb{Q}$, and although there is no general method to solve this problem, we do have a few useful tricks that can help in some cases.

Gauss's Lemma: Let $f \in \mathbb{Z}[t]$. If f is irreducible over $\mathbb{Z}$, then it is also irreducible over $\mathbb{Q}$.

This lemma is very useful because instead of having to consider factors with rational coefficients, we only need to consider factors having integer coefficients. This allows for divisibility arguments and other tricks.

Eisenstein's Criterion: Let $f(t) = a_0 + a_1 t + \cdots + a_n t^n \in \mathbb{Z}[t]$. Suppose there is a prime p such that

(1) $p \nmid a_n$ (2) $p | a_i$ for $0 \leq i < n$ (3) $p^2 \nmid a_0$

Then f is irreducible over $\mathbb{Q}$.

This criterion is proved using Gauss's Lemma and divisibility properties of primes.

Examples:

• $t^2 - 3, \quad 3t^8 - 2, \quad 2t^4 + 25t^3 - 15t^2 + 1000t + 45$ are irreducible over $\mathbb{Q}$.

• Eisenstein's Criterion could apply for more than one prime.

For example, $3x^7 - 10$ is irreducible over $\mathbb{Q}$.

Question 13: Prove that $\frac{2}{9}t^5 + \frac{5}{3}t^4 + t^3 + \frac{1}{3}$ is irreducible over $\mathbb{Q}$.

The next lemma follows from a clever application of Eisenstein's Criterion and the fact that if p is prime, then $p|\binom{p}{r}$ for $0 < r < p$.

Question 14: Prove that if p is prime, then $p|\binom{p}{r}$ for $0 < r < p$.

Lemma: If p is prime, then $f(t) = 1 + t + t^2 + \cdots + t^{p-1}$ is irreducible over $\mathbb{Q}$.

Question 15:

(a) Prove that given any positive integer n, there are irreducible polynomials in $\mathbb{Q}[t]$ with degree larger than n.

(b) By contrast, prove that if $f \in \mathbb{R}[t]$ is irreducible over $\mathbb{R}$, then $\partial f = 1$ or 2.

Hint for (b): Prove that if $\alpha \in \mathbb{C}$ is a zero of f, then so is its complex conjugate $\overline{\alpha}$. Deduce that the non-real zeros of f come in conjugate pairs.

The last irreducibility criterion has to do with reducing the coefficients of a polynomial modulo n.

Reduction Modulo n: There is a natural map $\mathbb{Z}[t] \to \mathbb{Z}_n[t]$ that reduces the coefficients modulo n.

For example, if $f = 3t^4 + 7t^3 + 6t^2 + 5t + 11 \in \mathbb{Z}[t]$, then

• Reduction modulo 2 maps f to $\overline{f} = t^4 + t^3 + t + 1 \in \mathbb{Z}_2[t]$

• Reduction modulo 5 maps f to $\overline{f} = 3t^4 + 2t^3 + t^2 + 1 \in \mathbb{Z}_5[t]$

Suppose that n does not divide the leading coefficient of f, so that the leading term does not disappear upon reduction modulo n, and so $\partial f = \partial \overline{f}$.

If $f = gh$, then from the laws of modular arithmetic we get $\overline{f} = \overline{g}\overline{h}$.

So, if f is reducible over $\mathbb{Z}$, then $\overline{f}$ is reducible over $\mathbb{Z}_n$. Therefore:

If $\overline{f}$ is irreducible over $\mathbb{Z}_n$, then f is irreducible over $\mathbb{Z}$

(and therefore also over $\mathbb{Q}$ by Gauss's Lemma)

This is very nice, because $\mathbb{Z}_n$ has only finitely many elements, so there are only finitely many possible factors of $\overline{f}$, and you can check them all.

The trick is to find a good value of n. In practice, you can usually choose it to be a prime p.

Question 16 (Challenge): Let a_0, a_1, a_2, a_3 and a_4 be odd integers. Must $f(t) = a_4t^4 + a_3t^3 + a_2t^2 + a_1t + a_0$ be irreducible over $\mathbb{Q}$?

Field Extensions

Definition: A field extension is a pair of fields K, L such that $K \subseteq L$.

We use the notation $L : K$.

Another way of saying this is to say that K is a subfield of L. A subfield is a subset of a field that is itself a field, and therefore it is closed under field operations: addition, multiplication, additive and multiplicative inverses.

The fields that we will work with are all subfields of $\mathbb{C}$. Note that any subfield of $\mathbb{C}$ must contain 0 and 1, and therefore (since it is closed under field operations) it must contain $\mathbb{Q}$.

Definition: Let K be a field and $X \subseteq K$ a subset. The subfield of K generated by X is the intersection of all the subfields of K that contain X.

Equivalently, it is the (unique) smallest subfield of K containing X.

If X contains a nonzero element, this is also equivalent to the set of all elements of K that can be obtained from elements of X by a finite sequence of field operations (addition, multiplication, additive inverses and multiplicative inverses).

We use the notation $\mathbb{Q}(X)$ for the subfield of $\mathbb{C}$ generated by X.

Definition: Let $L : K$ be a field extension and $Y \subseteq L$ a subset of the large field. The subfield of L generated by $K \cup Y$ is denoted by $K(Y)$.

When $Y = \{\alpha\}$ or $Y = \{\alpha_1, \dots, \alpha_n\}$, we write $K(\alpha)$ and $K(\alpha_1, \dots, \alpha_n)$ respectively, instead of $K(\{\alpha\})$ or $K(\{\alpha_1, \dots, \alpha_n\})$.

Examples:

• $\mathbb{Q}(i) = \{p + qi \mid p, q \in \mathbb{Q}\}$

• $\mathbb{Q}(\sqrt{2}) = \{p + q\sqrt{2} \mid p, q \in \mathbb{Q}\}$

• $\mathbb{R}(i) = \mathbb{C}$

• Let $\alpha = 2^{1/3} \in \mathbb{R}$. Then $\mathbb{Q}(\alpha) = \{p + q\alpha + r\alpha^2 \mid p, q, r \in \mathbb{Q}\}$

• $\mathbb{Q}(i, \sqrt{5}) = \{p + qi + r\sqrt{5} + si\sqrt{5} \mid p, q, r, s \in \mathbb{Q}\}$.

It takes some work to verify the last two examples. The third one is obvious. As for the first two,

Question 17: Prove that $L = \{p + qi \mid p, q \in \mathbb{Q}\}$ and $M = \{p + q\sqrt{2} \mid p, q \in \mathbb{Q}\}$ are in fact fields.

Hint: the only matter that is not completely straight forward is that these sets are closed under multiplicative inverses. That is, if $0 \neq z \in L$, then $z^{-1} \in L$, and similarly for M.

Definition: A simple extension is a field extension $L: K$ such that $L = K(\alpha)$ for some $\alpha \in L$.

That is, L is obtained from K by adjoining a single element α. Such an element is called a primitive element.

Obviously, the extensions $\mathbb{Q}(i)$, $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(2^{1/3})$ of $\mathbb{Q}$ that we saw in the examples above are simple extensions.

What is not so obvious is that the extension $\mathbb{Q}(i, \sqrt{5})$: $\mathbb{Q}$ is also a simple extension. In fact, $\mathbb{Q}(i, \sqrt{5}) = \mathbb{Q}(i + \sqrt{5})$. **Can you prove this?**

Definition: A field extension $L: K$ is finitely generated if there exist finitely many elements $\alpha_1, \dots, \alpha_n \in L$ such that $L = K(\alpha_1, \dots, \alpha_n)$.

It is clear that $K(\alpha_1, \alpha_2) = L(\alpha_2)$ where $L = K(\alpha_1)$, so an easy induction shows that any finitely generated field extension can be obtained by a finite sequence of simple extensions.

Algebraic and Transcendental Extensions

There is a crucial distinction between two kinds of simple extensions $K(\alpha): K$, which depends on whether or not α satisfies a nonzero polynomial over K:

Definition: Let $L: K$ be a field extension. An element $\alpha \in L$ is algebraic over K if $\exists\ 0 \neq f(t) \in K[t]$ such that $f(\alpha) = 0$. Otherwise α is transcendental over K.

If α is algebraic over K, we say that $K(\alpha): K$ is a simple algebraic extension.

If α is transcendental over K, we say that $K(\alpha): K$ is a simple transcendental extension.

Examples:

• The complex numbers $i, \sqrt{3}$ and $\alpha = 2^{1/3} \in \mathbb{R}$ are algebraic over $\mathbb{Q}$. They satisfy the nonzero polynomials $t^2 + 1, t^2 - 3$ and $t^3 - 2$ (which have coefficients in $\mathbb{Q}$) respectively. So $\mathbb{Q}(i), \mathbb{Q}(\sqrt{3})$, and $\mathbb{Q}(\alpha)$ are simple algebraic extensions of $\mathbb{Q}$.

• It is known that e and π are transcendental over $\mathbb{Q}$. These facts are not easy to prove. This means that, for example, π is not a root of any (nonzero) polynomial with rational coefficients. So $\mathbb{Q}(\pi): \mathbb{Q}$ is a simple transcendental extension.

Question 18: Prove that $\mathbb{Q}(\sqrt{3} + \sqrt{5}): \mathbb{Q}$ is a simple algebraic extension.

Notes:

• When a complex number $\alpha \in \mathbb{C}$ is algebraic (respectively, transcendental) over $\mathbb{Q}$, we simply say that α is algebraic (respectively, transcendental). That is, in this context, we drop the "over $\mathbb{Q}$". For example, $\sqrt{3}+\sqrt{5}$ is algebraic, and π is transcendental.

• It turns out that the set of all algebraic numbers forms a field. That means that this set includes 0 and 1 (**why?**) and is closed under field operations (addition, multiplication, additive and multiplicative inverses). The field of all algebraic numbers is denoted by $\mathbb{A}$. It is a subfield of $\mathbb{C}$.

Definition: The field extension $L:K$ is algebraic if every element of L is algebraic over K.

The Minimal Polynomial

Definition: A polynomial is monic if its leading coefficient is 1.

So, a monic polynomial has form $f(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$.

Note that the product of two monic polynomials is monic. Also, given a nonzero polynomial over a field, you can multiply it by a (unique) constant (the multiplicative inverse of the leading coefficient) and obtain a monic polynomial.

Claim: Let $L:K$ be a field extension and let $\alpha \in L$ be algebraic over K. Then there exists a unique monic polynomial $m(t) \in K[t]$ of smallest degree such that $m(\alpha) = 0$.

Question 19: Can you prove this claim?

Hint: By well order, there exists a monic polynomial $m(t) \in K[t]$ of smallest possible degree such that $m(\alpha) = 0$. Suppose that $m^*(t)$ is another such polynomial. Show that in fact $m = m^*$ by considering $m - m^*$.

Definition: Let $L:K$ be a field extension and let $\alpha \in L$ be algebraic over K. The unique monic polynomial $m(t) \in K[t]$ of smallest degree such that $m(\alpha) = 0$ is called the minimal polynomial of α over K.

This polynomial is also called the irreducible polynomial for α over K and is denoted it by $\text{irr}(\alpha, K)$. This is ok, because an alternative definition for it is as the unique monic *irreducible* polynomial $m(t) \in K[t]$ such that $m(\alpha) = 0$:

Lemma: Let $L:K$ be a field extension and let $\alpha \in L$ be algebraic over K. The minimal polynomial of α over K is irreducible over K, and it divides every polynomial $f(t) \in K[t]$ for which $f(\alpha) = 0$.

Let $m(t)$ be the minimal polynomial of α over K. Note of course that if $m \mid f$ in $K[t]$, then $f(\alpha) = 0$. Together with this lemma, we have that $m \mid f \Leftrightarrow f(\alpha) = 0$.

That is, the polynomials for which α is a zero are precisely the multiples of $m(t)$.

Examples:

• The minimal polynomial of i over $\mathbb{Q}$ or over $\mathbb{R}$ is $t^2 + 1$.

• The minimal polynomial of $\sqrt{3}$ over $\mathbb{Q}$ is $t^2 - 3$. However, the minimal polynomial of $\sqrt{3}$ over $\mathbb{R}$ is simply $t - \sqrt{3}$, since $\sqrt{3} \in \mathbb{R}$.

• More generally, if $\alpha \in K$, then the minimal polynomial of α over K is $t - \alpha$.

• Let $\xi_5 = e^{2\pi i/5}$, the 5th root of 1 that makes an angle of measure $2\pi/5$ (radians) with the positive x-axis. Clearly ξ_5 satisfies (is a zero of) the polynomial $t^5 - 1 \in \mathbb{Q}[t]$. Is this its minimal polynomial over $\mathbb{Q}$? No, because this polynomial is not irreducible over $\mathbb{Q}$. Clearly 1 is a zero, so $t - 1$ is a factor. It is easy to check that $t^5 - 1 = (t - 1)(t^4 + t^3 + t^2 + t + 1)$. Therefore, ξ_5 is a zero of $t^4 + t^3 + t^2 + t + 1$. This polynomial *is* the minimal polynomial of ξ_5 over $\mathbb{Q}$.

Question 20: Prove the last statement.

Vector Spaces

At this point we need to introduce some important concepts from linear algebra. You may already know a little about this subject if you studied systems of linear equations, matrices, and vectors.

Definition: A vector space V over a field K consists of a set V, a field K, and maps $+: V \times V \to V$ (vector addition) and $*: K \times V \to V$ (scalar multiplication) such that

(a) $(V, +)$ is an abelian group, and

(b) $\forall\, k, k_1, k_2 \in K$ and $\forall\, v, v_1, v_2 \in V$, scalar multiplication satisfies

- $k_1(k_2 v) = (k_1 k_2)v$
- $(k_1 + k_2)v = k_1 v + k_2 v$
- $k(v_1 + v_2) = kv_1 + kv_2$
- $1v = v$

It is easy to check that if V is a vector space over K, then:

- $0v = 0 \quad \forall\, v \in V$. Here the 0 on the left is in K and the 0 on the right is in V.
- $k0 = 0 \quad \forall k \in K$. Here 0 is in V.
- $(-k)v = k(-v) = -(kv) \quad \forall k \in K, v \in V$.

Examples:

- $\mathbb{R}^n$, the set of ordered n-tuples $(a_1, \ldots, a_n)$ of real numbers, is a vector space over $\mathbb{R}$, where addition of vectors is by components, $(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n)$, and scalar multiplication is given by $k(a_1, \ldots, a_n) = (ka_1, \ldots, ka_n)$. We can replace $\mathbb{R}$ with any field K.

• $K[t]$, the ring of polynomials over a field K, is a vector space over K where addition of vectors is the ordinary addition of polynomials and scalar multiplication is also the ordinary multiplication of polynomials, one of them being just a constant $k \in K$.

• If $M = M_{m\times n}(\mathbb{R})$ denotes the set of all $m \times n$ matrices with real entries, then M is a vector space over $\mathbb{R}$. Addition is the usual matrix addition, and scalar multiplication consists of multiplying every entry of a matrix by the scalar. We can replace $\mathbb{R}$ with any field K.

• Let V be the set of all functions $f: \mathbb{R} \to \mathbb{R}$. Then V is a vector space over $\mathbb{R}$. Addition is the usual (pointwise) addition of functions, and scalar multiplication is given by $(kf)(x) = kf(x)$ for all $x \in \mathbb{R}$. We could also take V to consist only of the continuous functions $\mathbb{R} \to \mathbb{R}$.

Example: If $L: K$ is a field extension, then L is a vector space over K. Addition of vectors and scalar multiplication are simply addition and multiplication in L.

That is, if $\alpha, \beta \in L$, then $\alpha + \beta$ is simply addition in L. As for scalar multiplication, say $k \in K$ and $\alpha \in L$. Since $K \subseteq L$, $k \in L$ too, and $k\alpha$ is simply multiplication in L.

This is the only kind of vector space that we will use.

Let V be a vector space over K.

Definition: A subset $S \subseteq V$ spans V if every $v \in V$ can be written in the form $v = k_1 v_1 + \cdots + k_n v_n$ for some $n \geq 0$, $k_i \in K$, and $v_i \in S$.

The sum $\sum_{i=1}^{n} k_i v_i$ is called a linear combination of $v_1, \ldots, v_n$.

So, S spans V if every vector in V is a linear combination of vectors in S.

When S spans V, we also say that the vectors of S span V.

Definition: The vector space V is finite dimensional if there exists a finite set $S \subseteq V$ that spans V.

Definition: The vectors in a subset $S \subseteq V$ are linearly independent over K if whenever $k_1 v_1 + \cdots + k_n v_n = 0$ with $n \geq 1$, $k_i \in K$, and $v_i \in S$, we must have $k_1 = \cdots = k_n = 0$. Otherwise, they are linearly dependent over K.

So, a set of vectors is linearly independent over K if the only way to express the vector 0 as a linear combination of these vectors is by having all the scalar coefficients equal to zero.

If the vectors are linearly dependent, there is a linear combination of them which equals 0 having scalar coefficients that are not all zero.

Definition: A basis for V over K is a set $B \subseteq V$ whose vectors span V and are linearly independent over K.

Every vector of V can be written uniquely as a linear combination of basis elements with scalar coefficients.

Example: A basis for $\mathbb{R}^3$ over $\mathbb{R}$ is $\{(1,0,0),(0,1,0),(0,0,1)\}$. It is easy to check that every element of $\mathbb{R}^3$ can be written uniquely as a linear combination of these three vectors, with real coefficients. This is just like setting up coordinates.

This example easily generalizes to $\mathbb{R}^n$ in an obvious way.

Finally, we come to the definition of the dimension of a vector space over a field:

Definition: Let V be a vector space over K. The dimension of V over K, denoted by $\dim_K V$ (or simply by $\dim V$ when K is understood), is the number of elements in any basis for V over K.

Examples:

• The dimension of $\mathbb{R}^n$ over $\mathbb{R}$ is n.

• The dimension of $M_{m\times n}(K)$ over K is mn.

• $\dim_{\mathbb{R}}\mathbb{C} = 2$. A basis for $\mathbb{C}$ over $\mathbb{R}$ is $\{1, i\}$.

• However, $\mathbb{C}$ is infinite-dimensional over $\mathbb{Q}$ (as you are asked to prove below).

Notes:

• The fact that $\dim_K V$ is well defined, i.e., that every vector space has a basis, and that any basis has the same number of elements, is proved in linear algebra courses.

• $\dim_K V$ is actually a cardinal number, meaning it is the cardinality of a set (a basis). It could be finite or infinite. If it is finite, that is, if V is finite dimensional, we write $\dim_K V < \infty$.

• Every set $S \subseteq V$ that spans V contains a subset $B \subseteq S$ that is a basis for V over K.

• Every set $S \subseteq V$ whose vectors are linearly independent over K can be enlarged to a basis B for V over K. By "enlarged" we mean $S \subseteq B$, of course.

Question 21: Prove that there does not exist a finite basis for $\mathbb{C}$ over $\mathbb{Q}$.

Hint: Find an infinite set of complex numbers that are linearly independent over $\mathbb{Q}$. Use the fact that π is transcendental.

The Degree of a Field Extension

Finally, we are ready to introduce the key concept that we will use to find a useful property of constructible numbers.

Also, now we start numbering some of our results, so that we can refer back to them in subsequent proofs.

Definition: The degree of the extension $L:K$, also called the degree of L over K, is the dimension of L as a vector space over K.

We denote the degree of $L:K$ by $[L:K]$. So $[L:K] = \dim_K L$.

If $[L:K] < \infty$, we say that the extension $L:K$ is finite.

Theorem 1: Let $K(\alpha):K$ be a simple extension.

If it is transcendental, then $[K(\alpha):K] = \infty$.

If it is algebraic, then $[K(\alpha):K] = \partial m$, where m is the minimal polynomial of α over K.

In fact, when α is algebraic over K, we have that $\{1, \alpha, \dots, \alpha^{n-1}\}$ is a basis for $K(\alpha)$ over K, where $n = \partial m$.

In this case, $[K(\alpha):K] = \partial m$ is also called the degree of α over K and is denoted by $\deg(\alpha, K)$.

Tower Law:

(a) If $K \subseteq L \subseteq M$ are fields, then $[M:K] = [M:L][L:K]$.

(b) More generally, if $K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n$ are fields then

$$[K_n:K_0] = [K_n:K_{n-1}][K_{n-1}:K_{n-2}] \cdots [K_1:K_0]$$

This law holds regardless of whether the extensions are finite or infinite, although we will only work with finite extensions. Part (b) follows from part (a) by an easy induction, and part (a) follows from the fact (**can you prove it?**) that if $A = \{x_i \mid i \in I\}$ is a basis for L over K and $B = \{y_j \mid j \in J\}$ is a basis for M over L, then $C = \{x_i y_j \mid i \in I,\ j \in J\}$ is a basis for M over K.

It is not hard to see, using the Tower Law, Theorem 1, and the quadratic formula, that:

Claim 1: Let $L\colon K$ be a field extension of degree 2, where $\operatorname{char} K \neq 2$ (this means that $2 = 1 + 1 \neq 0$ in K). Then $L = K\left(\sqrt{\alpha}\right)$ for some $\alpha \in K$.

The Constructible Numbers Revisited

Recall that the set $\mathfrak{C} = \{z \in \mathbb{C} \mid z \text{ is constructible}\}$ consists of those complex numbers z such that there is a finite sequence of ruler and compass constructions using R, C, P_{ll}, P_{lc} and P_{cc} that begins with 0 and 1 and ends with z.

Theorem 2: The set $\mathfrak{C}$ of all constructible numbers is a field. Moreover,

(a) $z = a + ib \in \mathfrak{C}$ if and only if $a \in \mathfrak{C}$ and $b \in \mathfrak{C}$. (Here $a, b \in \mathbb{R}$, of course)

(b) $z \in \mathfrak{C} \Longrightarrow \sqrt{z} \in \mathfrak{C}$.

(Of course, every nonzero complex number has two square roots. Statement (b) says that if z is constructible, then both of its square roots are constructible, since they are simply opposites, and $\mathfrak{C}$ is a field).

Denote by $L(z_1, z_2)$ the line that passes through z_1 and z_2 (here $z_1 \neq z_2$).

Denote by $C(z, r)$ the circle with center $z \in \mathbb{C}$ and radius $r > 0$.

Proof: To show that $\mathfrak{C}$ is a field, we must show that it contains 0 and 1, it is closed under addition and multiplication, it contains the opposite of any of its elements, and it contains the multiplicative inverse of any of its nonzero elements.

• By definition, $0 \in \mathfrak{C}$ and $1 \in \mathfrak{C}$.

• Given $0 \neq z \in \mathfrak{C}$, $L(0, z)$ and $C(0, |z|)$ intersect at $\pm z$. So $-z \in \mathfrak{C}$.

(Of course, if $z = 0$, then $-z = 0 \in \mathfrak{C}$)

• Given $z, w \in \mathfrak{C}$, then the "parallelogram law" tells us how to construct $z + w$ when $0, z$ and w are not collinear: $C(z, |w|)$ and $C(w, |z|)$ have $z + w$ as one of its points of intersection. When $0, z$ and w are collinear, it is even easier to construct $z + w$. So $z + w \in \mathfrak{C}$.

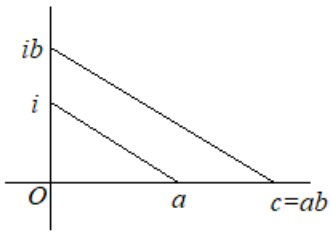
So far, we have proved that $\mathfrak{C}$ is a group under addition (and $1 \in \mathfrak{C}$).

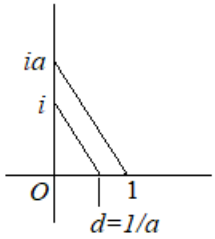
• Next we prove statement (a):

If $z = a + ib \in \mathfrak{C}$, drop perpendiculars from z to the x- and y-axes (which obviously can be constructed). We get $a \in \mathfrak{C}$ and $ib \in \mathfrak{C}$. Now $C(0, |ib|)$ intersects the x-axis at b, so $b \in \mathfrak{C}$.

Conversely, if $a \in \mathfrak{C}$ and $b \in \mathfrak{C}$ (where $a, b \in \mathbb{R}$), then $C(0, |b|)$ intersects the y-axis at ib. So $ib \in \mathfrak{C}$ and therefore $z = a + ib \in \mathfrak{C}$ by what we already proved.

• To prove that $\mathfrak{C}$ is closed under multiplication and reciprocals, we will first prove an intermediate result: $\mathfrak{C} \cap \{x \in \mathbb{R} \mid x > 0\}$ is closed under multiplication and reciprocals. The following picture shows why this is true. From a or b, we construct ia or ib, then draw appropriate parallel lines, and then we reason with similar triangles. You provide the details.





It follows immediately that $\mathfrak{C} \cap \mathbb{R}$ is a field.

Now we can show that $\mathfrak{C}$ is closed under multiplication and reciprocals:

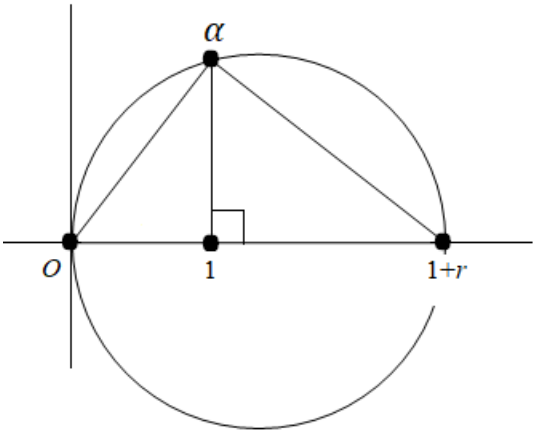
Let $z = a + ib \in \mathfrak{C}$ and $w = c + id \in \mathfrak{C}$. Then $zw = (ac - bd) + i(ad + bc)$. Using statement (a), then the fact that $\mathfrak{C} \cap \mathbb{R}$ is a field, and then statement (a) again, it follows that $zw \in \mathfrak{C}$.

Similarly, if $z \neq 0$, then $\frac{1}{z} = \frac{a}{a^2+b^2} - i\frac{b}{a^2+b^2}$, and using the same reasoning, we see that $\frac{1}{z} \in \mathfrak{C}$.

Now we have completed the proof that $\mathfrak{C}$ is a field.

• All that remains is to prove statement (b):

Let $0 \neq z \in \mathfrak{C}$. Write $z = re^{i\theta}$ with $0 < r = |z| \in \mathbb{R}$. We need to construct $w = \sqrt{r}e^{i\theta/2}$. It is clear since $z \in \mathfrak{C}$ that we can construct an angle of measure θ. Since we can bisect it, we can construct an angle of measure $\theta/2$. It is then also clear that if we can construct $\sqrt{r}$, then we can construct $w = \sqrt{z}$ and we will be done. You can check that it is possible to construct α as in the following picture, and (using similar triangles) that the segment joining 1 to α has length $\sqrt{r}$.



Therefore $\sqrt{z} \in \mathfrak{C}$ and we are done. □

Definition: The Pythagorean closure $\mathbb{Q}^{py}$ of $\mathbb{Q}$ is the smallest subfield $K \subseteq \mathbb{C}$ with the property that $z \in K \Longrightarrow \sqrt{z} \in K$.

That is, $\mathbb{Q}^{py}$ is the smallest subfield of $\mathbb{C}$ that contains the square roots of all its elements. It is the intersection of all subfields of $\mathbb{C}$ having that property.

It can also be described as the subfield of $\mathbb{C}$ obtained from 0 and 1 by using a finite sequence of field operations and square roots.

Theorem 3: $\mathfrak{C} = \mathbb{Q}^{py}$.

Proof: From the last theorem, $\mathfrak{C}$ is a subfield of $\mathbb{C}$ that contains the square roots of all its elements. By the definition of $\mathbb{Q}^{py}$ as the smallest such field, we have $\mathbb{Q}^{py} \subseteq \mathfrak{C}$.

The details of the reverse inclusion $\mathfrak{C} \subseteq \mathbb{Q}^{py}$ are tedious, but the idea is simple. Any $z \in \mathfrak{C}$ is obtained from 0 and 1 by a finite sequence of constructions using $\underline{\mathrm{R}}$, $\underline{\mathrm{C}}$, P_{ll}, P_{lc} and P_{cc}. Each constructed point along the sequence is the result of intersecting two lines, or a line and a circle, or two circles. And each such intersection is a solution of a system of equations that can always be expressed using field operations and square roots starting from previously constructed points (numbers). Therefore $z \in \mathbb{Q}^{py}$. □

Theorem 4: $z \in \mathfrak{C}$ if and only if there exists a finite tower

$$\mathbb{Q} = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n$$

of subfields of $\mathbb{C}$ such that $z \in K_n$ and $[K_i : K_{i-1}] = 2$ for $1 \leq i \leq n$.

Proof:

• First suppose that $\mathbb{Q} = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n$ is a tower of subfields of $\mathbb{C}$ such that $z \in K_n$ and $[K_i : K_{i-1}] = 2$ for $1 \leq i \leq n$. We show by induction that $K_i \subseteq \mathfrak{C}$ for $0 \leq i \leq n$. Then it follows that $z \in \mathfrak{C}$, since $z \in K_n \subseteq \mathfrak{C}$.

$\mathbb{Q} = K_0 \subseteq \mathfrak{C}$ because $\mathfrak{C}$ is a subfield of $\mathbb{C}$ and every subfield of $\mathbb{C}$ contains $\mathbb{Q}$.

Suppose $K_{i-1} \subseteq \mathfrak{C}$. Since $[K_i : K_{i-1}] = 2$, Claim 1 above says that $K_i = K_{i-1}(\sqrt{\alpha})$ for some $\alpha \in K_{i-1}$. Since $K_{i-1} \subseteq \mathfrak{C}$, $\alpha \in \mathfrak{C}$. Therefore $\sqrt{\alpha} \in \mathfrak{C}$, since $\mathfrak{C}$ contains the square roots of all its elements (recall either statement (b) of Theorem 2 or Theorem 3: $\mathfrak{C} = \mathbb{Q}^{py}$). Therefore $K_i = K_{i-1}(\sqrt{\alpha}) \subseteq \mathfrak{C}$. ▪

• Conversely, suppose that $z \in \mathfrak{C}$. Since $\mathfrak{C} = \mathbb{Q}^{py}$, z can be obtained from 0 and 1 (equivalently, from $\mathbb{Q}$) by a finite sequence of field operations and square roots. Each step in this sequence either can be performed in the same field or requires adjoining a square root that produces a field extension of degree 2. Therefore, discarding the repeated fields, we get a tower $\mathbb{Q} = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n$ of subfields of $\mathbb{C}$ such that $z \in K_n$ and $[K_i : K_{i-1}] = 2$ for $1 \leq i \leq n$. ▪ □

Finally, we are ready to state a useful necessary condition for a complex number to be constructible:

Theorem 5: If $\alpha \in \mathfrak{C}$, then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = \deg(\alpha, \mathbb{Q}) = 2^m$ for some $m \geq 0$.

Therefore, every constructible number is algebraic, and the degree of its minimal polynomial over $\mathbb{Q}$ is a power of 2, by Theorem 1.

Proof: Let $\alpha \in \mathfrak{C}$. By Theorem 4, there exists a finite tower

$$\mathbb{Q} = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n$$

of subfields of $\mathbb{C}$ such that $\alpha \in K_n$ and $[K_i : K_{i-1}] = 2$ for $1 \leq i \leq n$.

By the Tower Law, $[K_n : \mathbb{Q}] = 2^n$. But we also have $\mathbb{Q} \subseteq \mathbb{Q}(\alpha) \subseteq K_n$. Using the Tower Law again, $2^n = [K_n : \mathbb{Q}] = [K_n : \mathbb{Q}(\alpha)][\mathbb{Q}(\alpha) : \mathbb{Q}]$.

Therefore $[\mathbb{Q}(\alpha) : \mathbb{Q}] \mid 2^n$, and so $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^m$ for some $0 \leq m \leq n$. □

Caution: Theorem 5 gives a necessary condition for a complex number to be constructible. But it does not give a sufficient condition: its converse is false.

That is, if $\alpha \in \mathbb{C}$ has degree 2^m over $\mathbb{Q}$, it does not follow that $\alpha \in \mathfrak{C}$.

Question 22: Decide whether or not each of the following numbers can be constructed with ruler and compass:

(a) $\alpha = 2 + \sqrt{3 + \sqrt{5 + \sqrt{7}}} + i$

(b) The real root of the polynomial $f(t) = 5t^{11} + 10t^3 + 4t + 6$

Impossibility Proofs

Finally, we are ready to prove the impossibility of doubling the cube, trisecting the angle, and squaring the circle. Keep in mind, despite how easy this will be to us now, that these problems were open for about 2000 years!

Claim 2: $\alpha = 2^{1/3} \in \mathbb{R}$ (the real cube root of 2) is irrational.

Proof: Suppose for a contradiction that $\alpha \in \mathbb{Q}$. Write $\alpha = \frac{k}{l}$ where $k, l \in \mathbb{Z}$. Then $2 = \alpha^3 = \frac{k^3}{l^3}$, so $k^3 = 2l^3$. But this is impossible by unique prime factorization in $\mathbb{Z}$, because the power of 2 that exactly divides k^3 is a multiple of 3, while the power of 2 that exactly divides $2l^3$ is $\equiv 1 \bmod 3$. Therefore $\alpha \notin \mathbb{Q}$. □

Theorem 6: The cube cannot be doubled by ruler and compass.

Proof: Doubling the cube is clearly equivalent to constructing $\alpha = 2^{1/3} \in \mathbb{R}$.

But the minimal polynomial of α over $\mathbb{Q}$ is $t^3 - 2$. Therefore $[\mathbb{Q}(\alpha): \mathbb{Q}] = 3$.

By Theorem 5, $\alpha \notin \mathfrak{C}$. □

Question 23: Why is doubling the cube equivalent to constructing $\alpha = 2^{1/3} \in \mathbb{R}$?

For the impossibility of squaring the circle, we must take for granted the fact that π is transcendental (as we mentioned a while back, this is not easy to prove).

Theorem 7: The circle cannot be squared by ruler and compass.

Proof: Squaring the circle is clearly equivalent to constructing $\sqrt{\pi}$. Suppose that $\sqrt{\pi} \in \mathfrak{C}$. Then (since $\mathfrak{C}$ is a field) $\pi \in \mathfrak{C}$. But then Theorem 5 implies that π is algebraic, a contradiction. □

Question 24: Why is squaring the circle equivalent to constructing $\sqrt{\pi}$?

Theorem 8: The angle cannot be trisected by ruler and compass.

Proof: It suffices to exhibit a single angle that cannot be trisected. Below we exhibit such an angle. □

Of course, some angles can be trisected (like the straight angle and the right angle). But the theorem says that it is impossible to trisect a general, arbitrary angle.

Theorem 9: The angle of measure $2\pi/3$ cannot be trisected by ruler and compass.

Proof: Trisecting this angle is clearly equivalent to constructing $\xi_9 = e^{2\pi i/9}$. Now, if $\xi_9 \in \mathbb{C}$, then $\xi_9 + \xi_9^{-1} \in \mathbb{C}$. But we will show that $\alpha = \xi_9 + \xi_9^{-1} \notin \mathbb{C}$.

Let $w = \xi_3 = \xi_9^3 = e^{2\pi i/3}$. It is easy to see that $w^2 + w + 1 = 0$ (by direct calculation, or by symmetry, or by noting that the minimal polynomial of w over $\mathbb{Q}$ is $t^2 + t + 1$). Therefore $\xi_9^6 + \xi_9^3 = -1$. Now,

$$\alpha^3 = (\xi_9 + \xi_9^{-1})^3 = \xi_9^3 + 3\xi_9 + 3\xi_9^{-1} + \xi_9^{-3}$$

Since $\xi_9^{-3} = \xi_9^{-3} \cdot 1 = \xi_9^{-3} \cdot \xi_9^9 = \xi_9^6$, we get

$$\alpha^3 = \xi_9^3 + 3\xi_9 + 3\xi_9^{-1} + \xi_9^6 = 3\alpha - 1$$

and so α is a zero of the polynomial $t^3 - 3t + 1 \in \mathbb{Q}[t]$. But this polynomial is irreducible over $\mathbb{Q}$ (see below), so it is the minimal polynomial of α over $\mathbb{Q}$. Since its degree is not a power of 2, $\alpha \notin \mathbb{C}$ by Theorem 5. □

The irreducibility of $t^3 - 3t + 1$ over $\mathbb{Q}$ follows from Gauss's Lemma, since it is easy to see that it has no zeros in $\mathbb{Z}$ and therefore it is irreducible over $\mathbb{Z}$, by the Factor Theorem and the fact that the polynomial has degree 3. Even easier, it has no roots in $\mathbb{Q}$ by the "Rational Roots Theorem" from high school, which in this case tells us that the only possible rational roots are ± 1 (clearly neither is a root).

The Constructible Regular Polygons

We have proved the impossibility of doubling the cube, trisecting the angle, and squaring the circle by ruler and compass.

The other famous problem that the Greeks were unable to solve is to determine precisely which regular polygons can be constructed by ruler and compass.

The ancient Greeks knew how to construct the regular pentagon. Obviously, they also could construct the equilateral triangle, the square, and the regular hexagon.

In general, they were quite good at finding constructions when they are possible, but they were unable to prove the impossibility of other constructions. Also, they missed the construction of the regular 17-gon. This is understandable, as this construction is very elaborate. It was Gauss who found it (in 1796, when he was 19 years old). He went on to determine exactly which regular polygons can be constructed by ruler and compass. Here are some basic ideas:

• The regular n-gon is constructible if and only if $\xi_n = e^{2\pi i/n}$ is constructible.

We proved this near the beginning.

• Since we can bisect any angle, if we can construct the regular n-gon, then we can construct the regular $2n$-gon.

• If the regular m-gon and the regular n-gon are constructible and m, n are coprime, then the regular mn-gon is constructible.

This is true because, since $\gcd(m, n) = 1$, there exist integers a, b such that $am + bn = 1$. Therefore $a\frac{2\pi}{n} + b\frac{2\pi}{m} = \frac{2\pi}{mn}$.

Question 25: Finish the proof that the regular mn-gon is constructible if the regular m-gon and the regular n-gon are both constructible and $\gcd(m, n) = 1$.

The Greeks knew the last two bullets. Since they could construct the regular n-gon for $n = 3, 4, 5$, they knew that they could construct them for

$n = 3, 4, 5, 6, 8, 10, 12, 15, 16, 20, \ldots$

The first missing case in this list is $n = 7$. Can the regular heptagon be constructed with ruler and compass? No, it cannot:

Theorem 10: The regular heptagon cannot be constructed by ruler and compass.

Proof: We show, equivalently, that $\xi_7 = e^{2\pi i/7} \notin \mathbb{C}$. Since $\xi_7^7 = 1$, the minimal polynomial of ξ_7 over $\mathbb{Q}$ divides $t^7 - 1$.

But $t^7 - 1 = (t-1)(t^6 + t^5 + t^4 + t^3 + t^2 + t + 1)$ and the second factor is irreducible over $\mathbb{Q}$ (**why?**). Therefore, this second factor is the minimal polynomial of ξ_7 over $\mathbb{Q}$. Its degree is not a power of 2, so $\xi_7 \notin \mathbb{C}$ by Theorem 5. □

By Theorem 9, the regular 9-gon cannot be constructed either. How about the cases $n = 11, 13, 14, 17, 18, 19, \ldots$? Using the third bullet above, we can reduce the problem to the case when n is a prime power. But this is still a difficult problem.

Using marvelous ideas involving the interplay between group theory and field extensions, Gauss was able to completely settle the question:

Theorem 11: Let $n > 2$ be an integer. The regular n-gon can be constructed with ruler and compass if and only if $n = 2^r p_1 \cdots p_s$, where $r \geq 0$ and $s \geq 0$ are integers and the p_i are distinct Fermat primes.

A Fermat prime is a prime of form $p = 2^{2^k} + 1$. Fermat thought that $F_k = 2^{2^k} + 1$ is prime for all $k \geq 0$ ($F_0 = 3, F_1 = 5, F_2 = 17, F_3 = 257, F_4 = 65537$ are indeed prime). But he was wrong: Euler showed that F_5 is not prime. In fact, Fermat was very wrong: as of 2021, the above numbers 3, 5, 17, 257 and 65537 are the only known Fermat primes, and heuristic arguments suggest that these are very likely the only ones.

Question 26: Decide whether or not the regular n-gon is constructible from ruler and compass for each of the following values of n:

(a) 25 (b) 51 (c) 768 (d) 771 (e) p^2 for an odd prime p (f) 2^k for $k > 1$